Optimization and Evaluation of Gabor feature sets for ASR

Bernd T. Meyer and Birger Kollmeier

Medical Physics, Institute of Physics, University of Oldenburg

bernd.meyer@uni-oldenburg.de

Abstract

In order to enhance automatic speech recognition performance in adverse conditions, Gabor features motivated by physiological measurements in the primary auditory cortex were optimized and evaluated. In the Aurora 2 experimental setup such localized, spectro-temporal filters combined with a Tandem system yield robust performance with a feature set size of 30. Improved results can be obtained when using a Hanning window instead of a cut-off Gaussian envelope due to better modulation frequency characteristics. An analysis of complementarity of Gabor and MFCC features shows that errors could be reduced by 55% with a perfect classifier. In a real world scenario, a relative WER reduction of 15% compared to a competitive baseline is achieved by combining the feature types, indicating the potential of this class of physiologically motivated features.

Index Terms: spectro temporal features, automatic speech recognition, Gabor features

1. Introduction

The large gap in performance between human speech recognition (HSR) and state-of-the art automatic speech recognition (ASR) is most drastically encountered in adverse acoustic conditions and prohibits ASR technology from being widely used. Consistently, humans outperform machines by at least an order of magnitude [1]. In recent studies, the gap between human and automatic recognizers was found to be somewhat smaller, but error rates are still more than 150% higher for ASR than for HSR for a simple phoneme recognition task [2]. Human listeners recognize speech even in very adverse acoustical environments with strong reverberation and interfering sound sources. Our work is thus led by the idea of learning certain feature extraction techniques from the biological blueprint.

Findings from a number of physiological experiments in different mammal species showed that a large percentage of neurons in the primary auditory cortex (A1) respond differently to upward- versus downward-moving ripples in the spectrogram of the input [3]. Individual neurons are sensitive to specific spectro-temporal modulation frequencies in the incoming sound signal. Response patterns derived from STRFs were shown to correlate with articulatory features of phonemes (such as voicing or place or articulation) and result in confusion matrices similar to confusions from human listeners when used as features for ASR [4]. In the visual cortex, STRFs are measured with (moving) orientated grating stimuli. The results match very well two-dimensional Gabor functions. The use of 2D complex Gabor filters as features for ASR has been proposed earlier and proven to be relatively robust as part of a high end system [5]. This approach of spectro-temporal processing by using localized sinusoids most closely matches the neurobiological data and also incorporates other features as special cases: purely spectral Gabor functions perform an analysis similar to Mel-frequency cepstral coefficients (MFCCs)—modulo the windowing function—and purely temporal ones can resemble TRAPS or the RASTA impulse response and its derivatives [6] in terms of temporal extent and filter shape. Our approach is similar to the experiments presented in [7] where a 2D discrete cosine transform (DCT) was applied to patches of the short-time fourier transform. By using only lower coefficients as basis for ASR features, relevant information including spectro-temporal patterns can be extracted from speech. This feature extraction scheme proved to perform better than the well-known MFCC-HMM system, but uses relatively high dimensional features. The same is true for experiments with Gabor features, where typical feature vectors have 60 dimensions and additional delta and double-delta components [5].

In this paper, Gabor features are analyzed with respect to the required dimensionality and the optimal envelope shape. We report on complementary information of MFCCs and spectro-temporal features, and on the theoretical and practical improvements resulting from a combination of feature types.

2. Spectro-temporal Gabor features

A spectro-temporal representation of the input signal is processed by a number of 2-D modulation filters as shown in Fig. 1. The filtering is performed by correlation over time of each input frequency channel with the corresponding part of the Gabor function (centered on the current frame and desired frequency channel) and a subsequent summation over frequency. This yields one output value per frame per filter and is equivalent to a 2-D correlation of the input representation with the complete filter function and a subsequent selection of the desired frequency channel of the output. In this study, log mel-spectrograms serve as input features for feature extraction. This was chosen for its widespread use in ASR and because the logarithmic compression and mel-frequency scale might be considered a very simple model of peripheral auditory processing. The two-dimensional complex Gabor function $G(n, k)$ as proposed in [8] for ASR is defined as the product of a Gaussian envelope

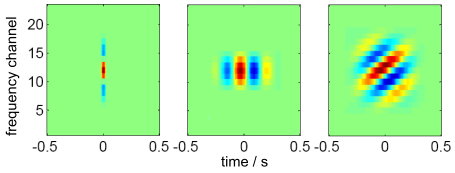

Figure 1: Examples for Gabor filter functions. Parameters allow for choosing purely spectral or temporal modulation filters (left and middle panel, respectively) or spectro-temporal filters (right panel).

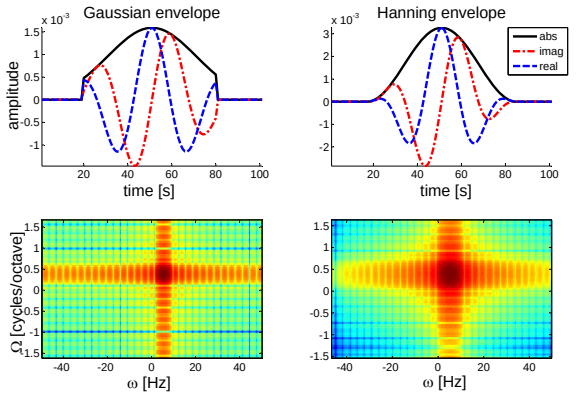

Figure 2: Illustration of 1-dim filter prototypes with cut-off Gaussian and Hanning envelope. The top row shows real and imaginary parts as well as envelopes of one dimensional Gabor filters, corresponding to a cross section of a 2-dim filter. The bottom row shows absolute values of spectro-temporal transfer functions of 2-dim Gabor filters (real values) plotted on logarithmic scale. The shading denotes amplitude in dB.

$g(n,k)$ and the complex sinusoidal function $s(n,k)$ (c.f. 1-D example in Fig. 2). The envelope width is defined by standard deviation values σ_n and σ_k, while the periodicity is defined by the radian frequencies ω_n and ω_k with n and k denoting the time and frequency index, respectively. The two independent parameters ω_n and ω_k allow the Gabor function to be tuned to particular directions of spectro-temporal modulation, including *diagonal* modulations. Further parameters are the centers of mass of the envelope in time and frequency n_0 and k_0. In this notation the Gaussian envelope $g(n,k)$ is defined as

$$g(n,k) = \frac{1}{2\pi\sigma_n\sigma_k} \cdot \exp\left[\frac{-(n-n_0)^2}{2\sigma_n^2} + \frac{-(k-k_0)^2}{2\sigma_k^2}\right]$$

and the complex sinusoid $s(n,k)$ as

$$s(n,k) = \exp\left[i\omega_n(n-n_0) + i\omega_k(k-k_0)\right].$$

The envelope width is chosen depending on the modulation frequency ω_x, respective the corresponding period T_x, either with a fixed ratio $\nu_x = T_x/2\sigma_x$ = 1 to obtain a 2D wavelet prototype or by allowing a certain range $\nu_x = 1..3$ with individual values for T_x being optimized in the automatic feature selection process. The infinite support of the Gaussian envelope is cut off at $1.5\sigma_x$ from the center. For time dependent features, n_0 is set to the current frame, leaving k_0, ω_k and ω_n as free parameters. From the complex results of the filter operation, real-valued features are obtained by using the real, imaginary or absolute part only. Special cases are temporal filters ($\omega_k = 0$) and spectral filters ($\omega_n = 0$). In these cases, σ_x replaces $\omega_x = 0$ as a free parameter, denoting the extent of the filter, perpendicular to its direction of modulation.

Alternatively, the filter can be designed as the product of a Hanning envelope $h(n,k)$

$$h(n,k) = 0.5 - 0.5\cdot\cos\left(\frac{2\pi(n-n_0)}{W_n+1}\right)\cdot\cos\left(\frac{2\pi(k-k_0)}{W_k+1}\right).$$

and the sinusoidal function $s(n,k)$, yielding the window lengths W_n and W_k as parameters instead of σ_n and σ_k.

3. Feature set optimization

In order to apply Gabor filters to the problem of speech recognition, parameter sets from a large number of possible combinations need to be determined. By putting further constraints on the spectro-temporal patterns, the number of free parameters can be decreased by several orders of magnitude. This is the case when a specific analytical function, such as the Gabor function [8], is explicitly demanded. Neurophysiological and psychoacoustical knowledge can be exploited for the choice of the prototype, as it is done here.

Feature set optimization is carried out by a modified version of a Feature-finding Neural Network (FFNN). It consists of a linear single-layer perceptron in conjunction with an optimization rule for the feature set . The linear classifier guarantees fast training, which is necessary because in this wrapper method for feature selection the importance of each feature is evaluated by the increase of RMS classification error after its removal from the set. This 'substitution rule' method [9] requires iterative retraining of the classifier and replacing the least relevant feature in the set with a randomly drawn new one. When the linear network is used for digit classification without frame-by-frame target labeling, temporal integration of features is carried out by simple summation of the feature vectors over the whole utterance, yielding one feature vector per utterance as required for the linear net. The FFNN approach has been successfully applied to digit recognition in combination with Gabor features in the past [8].

4. Experiments and results

4.1. Experimental setup

From American English digits strings (TIDigits corpus) and a set of Gabor filter prototypes, Gabor features were computed according to Section 2 and fed into a Tandem recognition system [10] as shown in Fig. 3. The N-dimensional feature vector was online normalized and combined with delta and double-delta derivatives before feeding into the multi-layer perceptron (MLP). The MLP was provided by the QuickNet software package (http://www.icsi.berkeley.edu) and had $3N$, 1000 and 56 neurons in input, hidden and output layer, respectively. It was trained on the TIMIT phone-labeled database with artificially added noise. The 56 output values were then decorrelated via principal component analysis (PCA, statistics derived on clean TIMIT) and fed into a hidden Markov model (HMM) trained on multicondition or clean speech data as proposed in the Aurora 2 experimental framework (see [11] for details). Experiments that aimed at the optimal number of features and optimization of the filter function were performed with the HTK reference recognizer. Analysis regarding complementarity of Gabor and traditional ASR features were carried out using a more competitive baseline: We used the Philips continuous ASR system [12] which is based on MFCCs (12 cepstral coefficients with Δ features which yields 24-dim. feature vectors) and an HMM classifier, and combines optimized feature extraction techniques such as non-linear spectral substraction (NSS) or noise-masking (NM) and classification based on discriminative training. Regarding training and test material, the same Aurora 2 signals as for the HTK system were used.

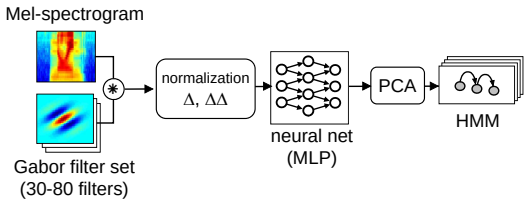

Figure 3: Schematic overview of the experimental setup. Feature vectors are obtained from correlation of Mel-spectrograms with Gabor filter prototypes and fed into a Tandem recognition system.

4.2. Optimal number of features

Higher number of features require more computation time and do not necessarily lead to improved recognition performance. In this experiment the number of feature components used as input for the Tandem system was varied from 10 to 80 features. Features were computed using the filter set G3 from [5] which was optimized on noisy German digits (ZIFKOM corpus). G3 yields improvements of over 50% compared to the baseline for clean training in a single stream experiment and improvements of 36% and 74% for noisy and clean training, respectively, in a multi-stream combination with the Qualcomm-ICSI-OGI front end. A reduction of number of features would result in fewer input neurons for the MLP, thus decreasing the total number of weights. For a fair comparison of classification performance, the number of neurons in the hidden layer was adjusted, so that the total number of weights remained constant at about 180k. The feature set contains 80 feature prototypes ordered by relevance. When using less than 80 features, the most relevant prototypes were chosen. In Fig. 4 the obtained error rates are shown. While WERs for multi condition training steadily decrease with higher number of features, this is not the case for clean condition training, where the error increases when using 80 instead of 70 features. However, both curves show saturation at 60 features, while performance superior to the baseline results is already achieved with 50 features for multi-condition training and 20 features for clean-condition training. The optimal number of features in the set would depend on application restrictions. Acceptable performance is reached with as few as 30 and optimal performance with 70 features for set G3. Note that delta and double-delta features would also be required, which would yield up to 240 features. The decrease of word accuracy from 70 to 80 features indicates that the least important 10 features in the set even have a detrimental affect on recognition performance, possibly a result of the optimization algorithm (c.f. Section 3).

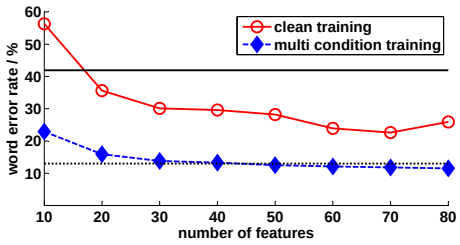

Figure 4: Averaged WERs depending on feature dimensionality: results are shown for clean and multi-condition training. MFCC baseline results are plotted as horizontal lines for multi condition training (dotted) and clean condition training (solid).

4.3. Envelope optimization

Cutting off the support of the Gaussian envelope at 1.5 σ results in unwanted higher harmonic frequencies in the modulation frequency domain. These distortions can be eliminated to a great extent by replacing the Gaussian envelope with a Hanning window. Fig. 2 shows a comparison of the spectro-temporal modulation transfer function of the two filter types.

In order to determine if the favorable modulation frequency characteristics of Hanning envelopes lead to improved recognition performance, eight filter sets with Gaussian and eight filter sets with Hanning envelope (each containing 60 filters) were generated by the automatic optimization procedure (Section 3) with ZIFKOM German digit data, mixed with different noise conditions. Temporal and spectral modulation frequencies were randomly chosen in an interval from 2 to 50 Hz and 0.06 to 0.5 cycles/octave, respectively. The width of the envelope was loosely coupled to the modulation frequency ω_x, using a value from 1 to 3 for the number of periods ν_x that lie in the interval $[-\sigma_x \;\; \sigma_x]$ for Gaussian envelopes or in the interval $[-W_x/1.5 \;\; W_x/1.5]$ for Hanning envelopes. Boundary conditions for ν_x guaranteed that even at low modulation frequencies the extension of the prototypes did not exceed 23 frequency channels or 101 time frames (corresponding to 1 second filter length). The results in Table 1 show that features based on Hanning-shaped envelopes outperform the baseline (13-dim. MFCCs with additional Δ and $\Delta\Delta$ features [11]) and features with Gaussian filter envelopes in all conditions. The best feature set with Hanning envelope HB02 also outperforms the reference feature set G3 and the best filter sets with Gaussian envelope.

		absolute WER			rel. improvement		
training condition		multi	clean	avg	multi	clean	avg
	a) baseline (MFCC)	13.0	41.9	27.5	0	0	0
Gabor-MLP	b) G3	12.1	23.9	18.0	6.9	43.0	24.9
	c) Avg Hanning	12.3	21.6	17.0	5.4	48.4	26.9
	d) Avg Gauss	13.2	23.7	18.5	-1.5	43.4	20.9
	e) Hanning HB02	12.0	19.5	15.7	7.7	53.5	30.6
	f) Gauss GB03	13.1	19.6	16.4	-0.8	53.2	26.2

Table 1: WERs for different Gabor filter sets. Beside the baseline data (a), results are shown for filter set G3 (b), averaged values for eight Hanning and eight Gaussian envelope sets (c & d) and best Hanning and Gaussian envelope sets (e & f).

4.4. Complementarity to common ASR features

While the limitation to purely spectral information is a theoretical disadvantage of MFCCs, it is often difficult to achieve improvements for tuned ASR systems with completely new features. We therefore investigated if both feature types carry complementary information and if a combination of Gabor and MFCC features is a promising approach. For these experiments, we chose results from the Philips continuous recognizer [12] with an improved feature extraction stage as baseline (see Section 4.1). The intersection of misclassified digit tokens E from both systems was chosen as a measure for complementary information: $I_{err} = (E_{Gabor} \cap E_{MFCC})$. The smaller I_{err} is, the smaller is the error rate of a (imaginary) perfect classifier that can use the MFCC or the Gabor feature information, and thus only produces an error if a digit was misclassified by *both* single-stream systems. A low error rate of such a perfect

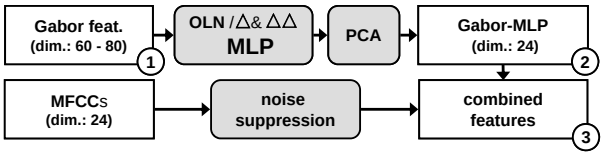

Figure 5: Feature combination scheme: Gabor features (1) are processed with an neural network (MLP). The output is decorrelated and its dimension is reduced to 24 with a PCA (2). Denoised MFCCs and the Gabor-MLP features are concatenated (3) and used as input to an HMM.

	absolute WER			relative improvement		
training condition	multi	clean	avg	multi	clean	avg
a) MFCCs	16.3	45.1	30.7	-89.3	-342.7	-216.0
b) Denoised MFCCs	8.6	10.2	9.4	0	0	0
c) Gabor-MLP	10.3	18.8	14.6	-20.0	-84.8	-52.4
d) Oracle	3.7	4.7	4.2	57.0	53.9	55.4
e) b + c	6.6	9.3	8.0	23.0	8.3	15.7

Table 2: WERs of HMMs fed with MFCC or Gabor-MLP features. Denoised MFCC features were chosen as baseline. Oracle WERs show the theoretical improvements that a perfect classifier could achieve. The best real-world performance is obtained with a stream combination of MFCCs and Gabor features processed by a MLP (inverted row).

or 'oracle' system represents a high complementarity of feature streams. The WERs of both feature types and the oracle system are shown in Table 2. While Gabor-MLPs alone cannot improve the baseline results, the perfect knowledge scenario decreases the error rates about 55% relative (row d).

This result motivated a combination of feature streams as depicted in Fig. 5: Denoised MFCCs are concatenated with Gabor features processed by the MLP and fed into a HMM. The results (Table 2 (e)) show that WERs can be reduced by 16% on average, and by 23% for multi-condition training. Compared to MFCCs without denoising, the average reduction in WER is over 70%. Gabor features were calculated with the filter set HB02 which gave better performance than G3, confirming the results from Section 4.3. However, this is not even halfway to the WER reduction observed for the oracle system, which motivates more advanced feature combination techniques such as LDA or the combination of multiple neural nets which will be subject to future work.

5. Discussion & conclusions

Results show that spectro-temporal features have advantages over conventional features in a standard ASR system, especially when the classifier is trained on clean speech. This emphasizes the robustness of Gabor features which could contribute to narrow the gap between HSR and ASR.

A drawback of Gabor features compared to traditional MFCCs is the higher number of required features which increases computational cost. We have shown that using the 20 most relevant features with delta and double-deltas already results in high performance superior to the Aurora 2 baseline for clean training. When using 50 feature components, performance exceeds the baseline for multi-condition training. Hanning-shaped Gabor filters show sharper modulation frequency characteristics and lead to increased performance compared with baseline results and feature sets with Gaussian envelope.

When feeding Gabor features directly to an HMM, we have so far not found this approach to improve performance over baseline which may be due to stronger model assumptions for the HMM compared to the MLP. Furthermore, Gabor features alone did not improve results of a system with advanced denoising techniques prior to MFCC calculation, which motivates modifications to the FFNN, e.g., by including MFCCs in the feature selection process. The Gabor filter set with best performance exhibits 30% purely spectral and temporal filters, respectively, while 40% of the automatically defined filters are spectro-temporal. An inclusion of MFCCs during the parameter definition process would presumably result in a shift towards spectro-temporal and purely temporal filters which would increase complementary information.

The applied filter sets were not designed with a combination with MFCCs in mind, but still result in an increase of performance when feature streams are concatenated. Relative WERs were reduced both in a theoretical approach (55%) as well as in a real-world scenario (16%), which demonstrates the potential of this class of physiologically motivated features.

6. Acknowledgments

Supported by the DFG (SFB/TR 31 'The active auditory system'; URL: http://www.uni-oldenburg.de/sfbtr31). We would like to thank Michael Kleinschmidt, David Gelbart, Thomas Brand and Alexander Fischer for their support and contribution to this work.

7. References

[1] R. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, pp. 1–15, 1997.

[2] B. Meyer, M. Wächter, T. Brand, and B. Kollmeier, "Phoneme confusions in human and automatic speech recognition," in *Proc. Interspeech*, 2007.

[3] D. Depireux, J. Simon, D. Klein, and S. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *J. Neurophysiol.*, vol. 85, pp. 1220–1234, 2001.

[4] N. Mesgarani, S. David, and S. Shamma, "Representation of phonemes in primary auditory cortex: How the brain analyzes speech," in *Proc. ICASSP*, 2007.

[5] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with Gabor feature extraction," in *Proc. ICSLP*, 2002.

[6] H. Hermansky, "Should recognizers have ears?" *Speech Communication*, vol. 25, pp. 3–24, 1998.

[7] T. Ezzat, J. Bouvrie, and T. Poggio, "Spectro-temporal analysis of speech using 2-d gabor filters," in *Proc. Interspeech*, 2007.

[8] M. Kleinschmidt, "Methods for capturing spectro-temporal modulations in automatic speech recognition," *Acustica united with acta acustica*, vol. 88, pp. 416–422, 2002.

[9] T. Gramß, "Fast algorithms to find invariant features for a word recognizing neural net," in *IEEE 2nd International Conference on Artificial Neural Networks*, Bournemouth, 1991, pp. 180–184.

[10] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, 2000.

[11] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *ISCA ITRW ASR*, 2000.

[12] M. Lieb and A. Fischer, "Progress with the philips continuous ASR system on the Aurora 2 noisy digits database," in *Proc. ICSLP*, 2002, pp. 449–452.